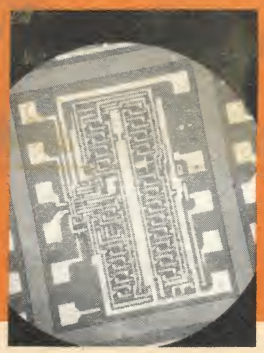
MICROELECTRONICS
Technical Bulletin



APPLICATION NOTES ■ TECHNICAL PAPERS ■ RELIABILITY REPORTS ■

Designing With MOS Semiconductors

by J. Leland Seely, Ph.D.
Manager, Integrated Circuits Engineering
Microelectronic Division
General Instrument Corp.

In the field of integrated circuits, industry leaders have often predicted the early arrival of a breakthrough technology capable of replacing today's triple-diffused microcircuits. The accuracy of those predictions has now been amply substantiated by the emergence and commercial availability of a new generation of devices known as MOS semiconductors. The author discusses here several basic considerations in the design of circuitry utilizing MOS devices – both discrete and integrated – recently introduced by General Instrument.

Introduction

It is just three years shy of two full decades since Bardeen and Brattain made their 1948 announcement on the development of the transistor, yet only recently has there emerged a semiconductor device capable of combining the high input impedance of the vacuum tube with all the advantages offered by the transistor (especially the ability to be d-c coupled without level shifting). This device, the Metal Oxide Silicon Transistor, or MOST*, not only provides higher input impedances ($\approx 10^{14}\Omega$), but also, because of the processing breakthrough it represents, opens up numerous opportunities for substantial progress in the fabrication of integrated circuits—opportunities which are at this moment resulting in lower cost and greater versatility for the equipment circuit designer.

In order to take full advantage of MOS devices —whether discrete or integrated—some knowledge of how these units operate is essential.

How the MOS Technology Achieves High Input Impedance

The name MOS is derived from the sandwich-like structure of Metal (usually aluminum), Oxide (SiO_2) and Silicon shown in Figure 1.

The metal forms the central electrode (gate) and is isolated from the body of the device by a thin but very non-conductive layer of silicon oxide. This oxide layer accounts for the extremely

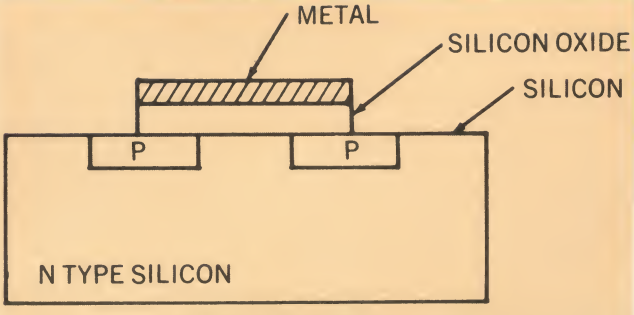


Figure 1. MOST Structure.

*Other names for this type of device are: Insulated Gate Field Effect Transistor (IGFET), and Metal Oxide Silicon Field Effect Transistor (MOSFET).

high input resistance ($\approx 10^{14}\Omega$) of the device.

Diffused into the basic N type silicon substrate are two P regions designated "source" and "drain". The area directly below the gate and lying between the source and drain is the channel of the device. When there is no voltage applied to the gate, the channel is N type silicon since it is just part of the basic substrate. Consequently, the source, channel, and drain form a PNP structure which looks electrically like a pair of back-to-back diodes. If a potential difference is applied between source and drain, no current will flow because, regardless of polarity, one PN junction will always be reverse biased. If, however, the gate is made sufficiently negative with respect to the source, holes are drawn into the channel region causing it to change from N type to P type.

The gate voltage at which this inversion occurs is called the threshold voltage. Increasing the magnitude of the gate voltage beyond this point allows ohmic conduction from source to drain. The conductivity depends on the depth of the channel which in turn depends on the magnitude of the gate to source potential.

The MOST can thus be considered as a variable resistor which is controlled by the gate voltage. The resistance between source and drain remains very high ($\approx 10^{10}\Omega$) as the gate is made more and more negative until the threshold voltage of the device is reached. At this point, the resistance decreases rapidly.

The resistance continues to decrease with more negative gate voltage, but the rate of change diminishes and the resistance approaches a limiting value which is determined by the geometry of the device ($\approx 300\Omega$ for the General Instrument MEM511 MOSFET).

The threshold voltage is an important characteristic of the MOST. It is this property which gives digital circuits employing MOST's their extremely high noise immunity. Noise signals which would saturate regular bi-polar transistors are completely rejected by the MOST if they are below the threshold voltage, which is approximately –4 volts.

In summary, then, the MOST has higher input resistance than the best vacuum tubes but can be d.c. coupled without need of level shifting. It has much higher noise immunity than bi-polar transistors, but still maintains the advantages of low power, small physical size, rugged construction and ability to operate without filiments.

MOS Device Applications

Because of its ideal characteristics, the MOST has applications throughout the entire spectrum of the electronics field. This has made it necessary initially to focus attention on some restricted segments of the industry in which large volume MOST applications are immediately feasible. For reasons which will become apparent later in this discussion, switching applications have been chosen – specifically, chopping, multiplexing and digital logic.

The use of FET's as choppers is now very common because there is no "offset" voltage associated with a field effect device. This is also true with a MOST (see Fig. 2A), since it is

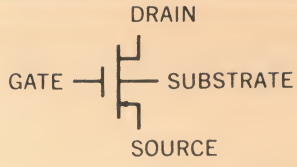


Figure 2A. Schematic Symbol For MOST. (Note: Substrate lead will not be shown in schematics which follow.)

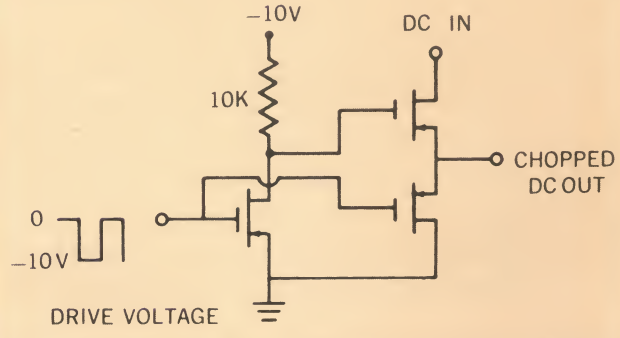


Figure 2B. Series-Shunt Chopper with Driver Circuit.

fundamentally a field effect transistor. Figure 2B shows a simple series-shunt chopper, with an inverter stage used to provide the out-of-phase drive. In this circuit, and all subsequent circuits, substrate leads are not shown but are assumed to be connected in common and returned to the most positive potential associated with the circuit.

The same precautions concerning capacitive feed-through spikes must be observed with MOST's as with conventional FET's. There is a direct trade-off available to the designer between feed-through capacitance and saturation resistance. This trade-off is strictly the result of geometry and comes about because the saturation resistance is inversely proportional to the channel length, and therefore to gate length. Gate capacitance, on the other hand, is directly proportional to gate length. (The gate metal and the substrate separated by a dielectric layer form a parallel plate capacitor). The impedance level of the input signal determines the proper choice of MOST.

Figure 3. gives an example of the use of MOST's to switch any one of four inputs to one

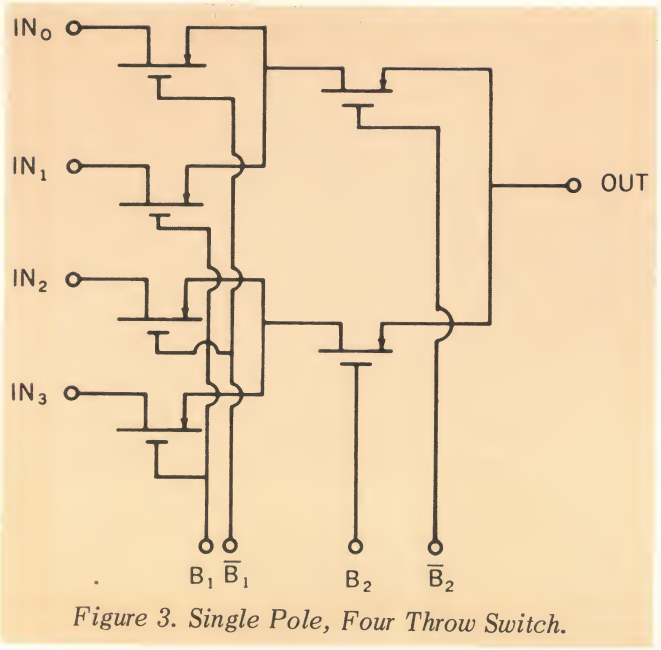


Figure 3. Single Pole, Four Throw Switch.

output. The switch is driven by inputs B_1 and B_2 and their complements $\overline{B}_1$ and $\overline{B}_2$.

By defining the "1" state as –10 volts and the "0" as zero volts, B_1 and B_2 form a binary number that tells which input is connected to the output. As the function to be performed becomes more complicated, the number of devices required to perform the function increases, and it soon becomes obvious that many advantages would accrue from putting interconnected arrays of devices into a single package, i.e., make integrated circuits.

The relative ease with which this can be done constitutes far and away the most important attribute of the MOST. To gain proper appreciation of this fact, it is necessary to consider the process by which a single MOST is made, and how the exact same process can be used to make a complete monolithic MOS integrated circuit.

MOS Fabrication

There are four basic steps involved in the fabrication of a MOST, each step involving use of a photographic mask. Fabrication begins with an N type silicon wafer which has had a thick layer of SiO_2 grown on its surface. The surface is coated with a photo resist material and exposed through mask #1. Mask #1 consists of a single rectangle so that, after developing, photo resist covers the entire surface except for this rectangular region. The wafer is next etched in an acid which attacks SiO_2 but does not attack the photo resist or Si. After etching, the resulting wafer has a single rectangular hole which cuts through the SiO_2 and stops at the surface of the silicon substrate. This is shown in Figure 4.

In order to define the channel region and simultaneously provide the thin oxide layer upon which the gate electrode will be deposited, the wafer is placed in a controlled ambient furnace and another thin layer of SiO_2 is grown. (Figure 5.)

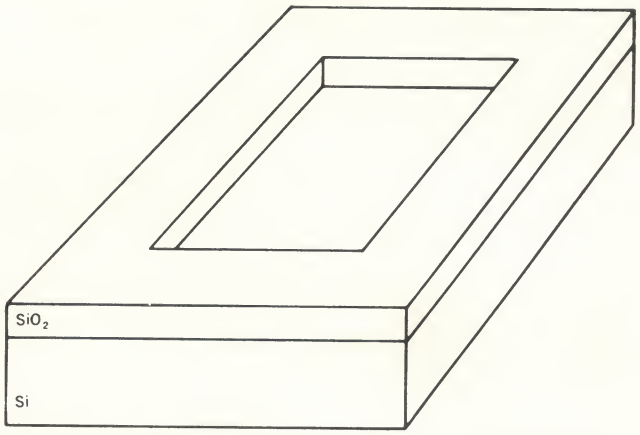


Figure 4A. Wafer After First Masking and Etching.

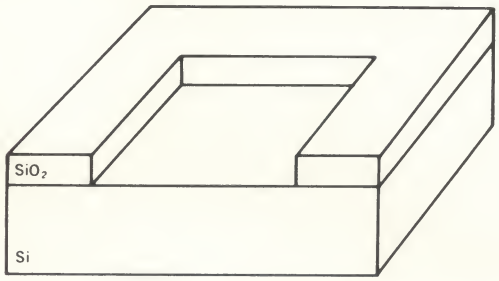


Figure 4B. Cross Section of Same Wafer.

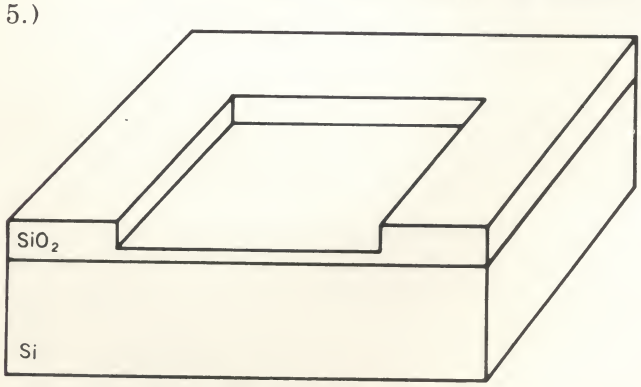


Figure 5. Cross Section of Wafer After Growth of Thin Oxide.

By use of Mask #2 and another etch, the thin oxide is left as shown in Figure 6.

The wafer is now ready for diffusion of the P regions which will form the source and drain of the device. The SiO_2 (both thick and thin regions) will block the diffusion of dopant into the silicon covered by these regions so that the P regions will be formed only under the areas of bare Si. Actually, some lateral diffusion takes place and the P regions extend slightly under the oxide.

For reasons which will be explained later, at the same time the P regions are being diffused,

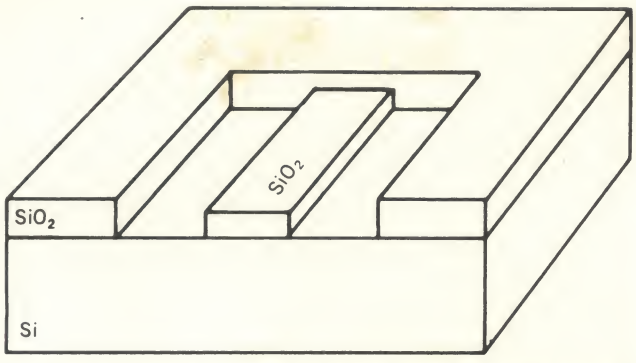


Figure 6. Cross Section of Wafer After Second Etch.

another thin layer of oxide is grown over the bare silicon. Since the rate of growth of oxide is exponentially dependent upon the thickness of oxide already present, the thick oxide grows very slightly. The thin oxide grows slowly and the new oxide grows rapidly. Proper control must be exercised in order to assure that the oxide under the gate is the proper thickness at the time the diffusion of the P regions is completed. At the end of this step, the structure is as shown in Figure 7.

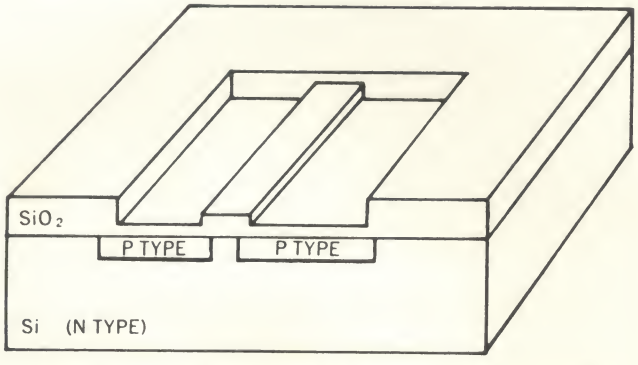


Figure 7. Cross Section of Wafer After Diffusion.

The third masking operation simply cuts slots in the thin oxide to permit ohmic contact to the two P regions. The final step is to vacuum evaporate aluminum over the entire surface of the wafer and then, with Mask #4, etch the metal into the desired pattern. This is shown in Figure 8. Metal from the contact holes is extended to the thick oxide region and formed into a bonding pad area for the subsequent nailhead bonding operation.

It is readily apparent that by merely changing the final metalization mask so that the contacts to the P regions and the extensions of the gates are interconnected (instead of being brought to separate bonding pads), a complete integrated circuit can be made.

One last problem remains – that of crossovers. In complicated circuits, two metal runs invariably need to pass over each other without shorting together. This can be accomplished without add-

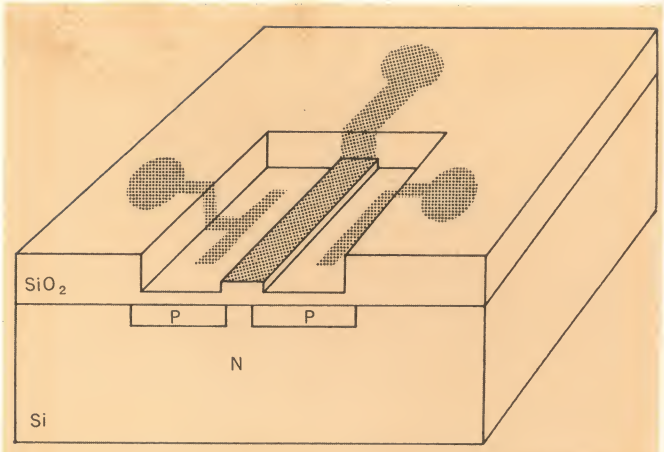


Figure 8. Cross Section of Completed MOST.

ing any additional steps to the process simply by providing an additional P region at the proper point. As the two metal runs approach each other, one stops and goes down through a contact hole in the oxide to a P region. The P region extends under and beyond the second metal run where it is contacted again by the first run. This is depicted in Figure 9. It was to allow for these

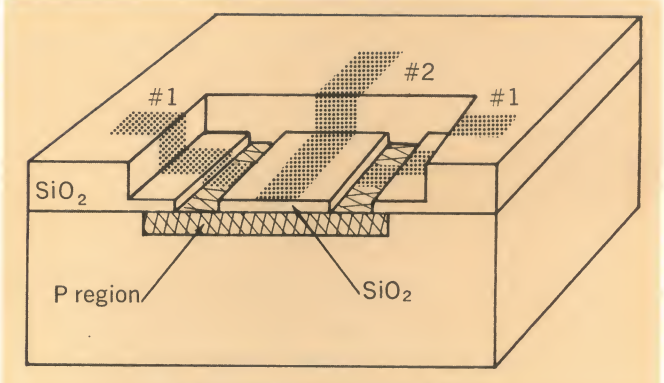


Figure 9. Crossover Point.

crossovers that the new oxide was grown over the P regions during the diffusion process.

Since the P region is not as good a conductor as the metal run, account must be taken of the resistance added to an interconnection when a crossover is used.

MOS Integrated Circuits

Again, it should be stressed that the only difference between making a MOS integrated circuit and making a single MOST is in the design of the masks. The processing is identical for both. This is extremely important in that it allows much faster new circuit development, since only the masks need be designed and no new diffusion schedules need to be worked out.

The simplicity of the MOS fabrication process is also important from an economic standpoint, since it provides higher yield of good circuits as well as the ability to make much more complex

circuits to an acceptable yield. By way of comparison, standard triple diffused integrated circuits usually have less than 20 or 30 transistors, while MOS integrated circuits commonly have 100 to 200 active devices (circuits have actually been made with 800 MOST's on a single monolithic chip).

The capability of making much more complicated circuits than heretofore possible can be exploited in several ways, with tradeoffs between cost, reliability and space. When space is at a premium, large amounts of circuitry should be put on a single chip. This has the attendant increase in reliability because of the decreased number of interconnections. It also increases cost per chip, and depending upon the amount of circuitry, may or may not increase the cost per circuit function. The field of integrated MOS circuits is still too new to say exactly where optimum circuit complexity is as far as cost goes, but it is definitely much beyond the complexity presently offered by triple diffused techniques.

Until now, a major design objective of digital integrated circuits has been to minimize the number of transistors; designing with MOST's, however, calls for minimizing the number of input and output leads, since generally more area is occupied by bonding pads and output stages than by all the rest of the circuitry. Output stages consume a large amount of area because the saturation resistance is inversely proportional to channel length. To achieve a low output impedance, therefore, a large geometry device must be incorporated at the output. Definite advantages accrue if an all-MOS system is used, in which case the output stages of various blocks of the system drive only other MOS devices.

When very low output impedance is required, a type of push-pull circuit is used. As shown in Figure 10, two MOST's, which are connected in series, are driven out of phase by an inverter stage. When V_i goes negative, say –12 volts, Q_1 and Q_3 are turned on.

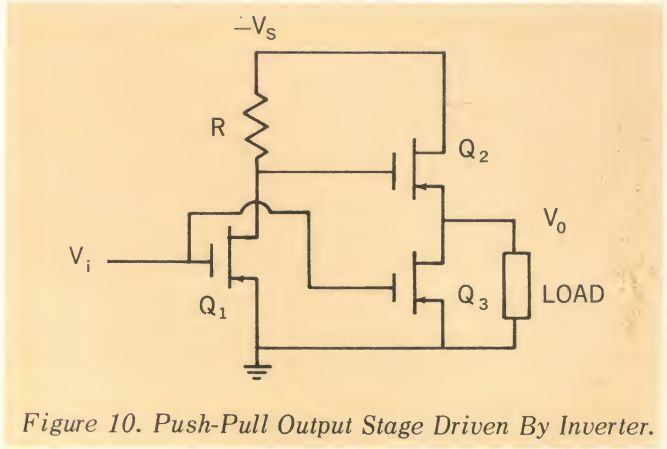


Figure 10. Push-Pull Output Stage Driven By Inverter.

With Q_1 on, the gate of Q_2 is at ground, which means Q_2 is off. V_o is thus returned to ground through the on resistance of Q_3. When V_i goes to ground, Q_1 and Q_3 turn off. The gate of Q_2 goes to $-V_s$ turning Q_2 on. V_o is thus returned to $-V_s$ through Q_2. Q_2, however, will not in general be in a saturated condition since it is operating in a "source follower" mode.

The impedance through Q_2 to $-V_s$ will depend on the magnitude of V_o, which in turn will depend upon the load being driven. As V_o approaches $-V_s$, the potential difference between the gate and source approaches the threshold voltage and starts to turn Q_2 off. This raises the impedance of Q_2, which allows V_o to be pulled towards ground by the load. But as V_o moves toward ground, the gate to source voltage of Q_2 increases, Q_2 is turned on harder, and V_o is pulled back towards $-V_s$. Thus there is a stable feedback effect which reduces output impedance to the negative supply as the output is more heavily loaded.

Nevertheless, for equal size MOST's, Q_3 will have a lower impedance to ground than Q_2 will have to $-V_s$, simply because of the difference in gate to source voltage. In addition, even under no load conditions, V_o can approach only to within one threshold voltage of $-V_s$. In this regard it should be noted that when the source is more negative than the substrate, as is the case when Q_2 is on, the effective turn-on voltage is more negative than the normal –4 volts.

How MOS Solves the Resistor Problem in Integrated Circuits

Nearly all integrated circuits require resistors. These are commonly made by either diffusion or by thin film techniques. There are problems associated with each type. Since it is desirable to keep the circuit area at a minimum, the use of small geometry MOST's is dictated. This means high saturation resistance, which in turn means load resistors must have large values, typically 100k ohms or more. If these resistors are to be made by diffusion without changing the basic fabrication process, they must have the same sheet resistivity as the P regions which constitute the source and drain, i.e., 50 ohms per square. Because of limitations on photo resist techniques, the width of a diffused line must be .0002 inches or greater. Thus a line .001 inches long can be no more than 5 squares, or 250 ohms. To make a 100k ohm resistor, then, requires a line .400 inches long. This is far too long to be of practical use in circuits requiring many resistors. Even if the process is changed to allow

a different diffusion schedule for the resistors, resisitivity values of greater than 500 ohms per square are difficult to control. This means the resistor will be at least .040 inches long, which is still objectionable.

Cermet thin film resistors can be made with 10,000 ohms per square, which would be highly acceptable area-wise were it not for a power dissipation problem. With a supply of −25 volts, the power dissipated by a 100K ohm resistor is 6mw. The area occupied by the resistor is 10 squares of dimension .0002 inches (or 4×10^{-7} square inches). This is 15,000 watts per square inch! Cermets don't hold up too well under these operating conditions. They do show promise in applications where more area is available but, again, additional process steps are required.

The best solution to the resistor problem at present is to use MOST's. By proper geometry design, the saturation resistance can be made as high as 200,000 ohms in an area only 1.6×10^{-6} square inches. Thus, by returning the gate of a small geometry MOST to the negative supply, a large value, small area resistor is created. Figure 11 shows a MOST used as a load resistor in an inverter stage.

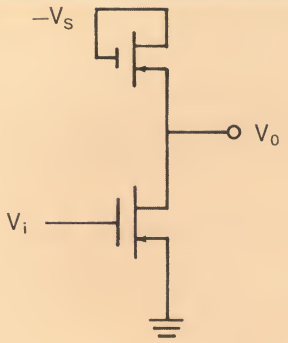


Figure 11. Inverter Stage Using Small Geometry MOST as a Load Resistor.

As pointed out in conjunction with the push-pull output stage of Figure 10, V_o can pull virtually to ground but can swing only to within one effective threshold voltage of the gate of Q_2. Thus, if R is really a small geometry MOST, V_o is *two* threshold voltages away from the supply. For a supply voltage of −25 Volts, the output can be expected to swing only to − (25-7-6) = −12 volts. (The effective threshold voltage of Q_2 will be less than that of the resistor MOST since the source to substrate voltage is less). This double threshold voltage drop accounts for the relatively high supply voltages used in MOS integrated circuits.

Another approach to the resistor problem is to return all the gates of the resistor MOST's to a separate supply voltage of, say, −25 volts and then reduce V_s to around −12 volts. Since the gates draw no current, this drops circuit power dissipation a factor of more than 4. The output stage can now be of the simple inverter type and V_o will swing all the way to $-V_s$. The savings in power can be spent to increase speed and lower output impedance if necessary. The ability to change values of load resistors throughout the circuit by means of an external supply proves to be extremely convenient at times.

MOS Microcircuit Applications

In the actual design of MOS integrated digital circuits, it is often possible to make use of dynamic steering by using charge storage capacitors. Because of the extremely low leakage of the associated MOST's, these capacitors can be a fraction of a picofarad and still have usefully long-time constants. Two examples of this usage are the R S T flip-flop shown in Figure 12 and the twenty-one bit shift register shown in Figure 13.

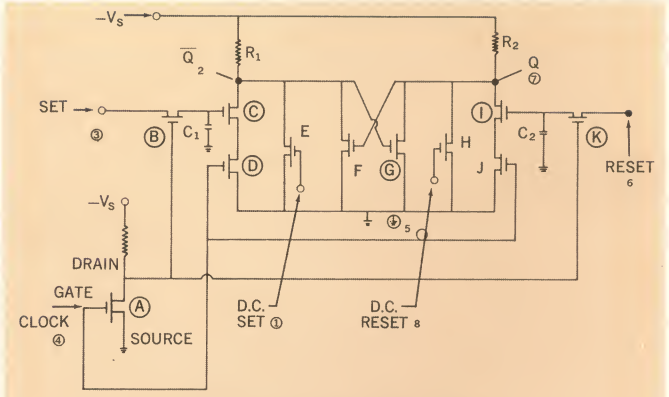


Figure 12. RST Flip Flop.

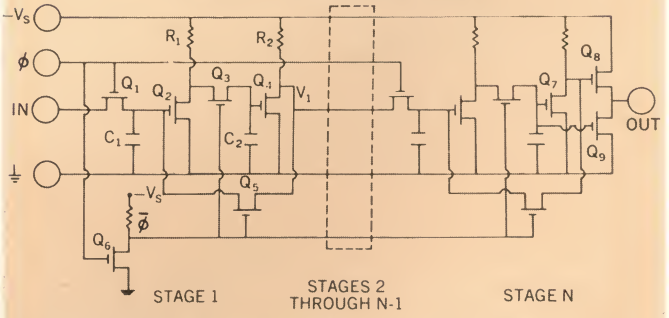


Figure 13. 21-Bit MOS Monolithic Shift Register.

In those cases where the entire circuit is integrated on a single chip of silicon, all substrate leads are of necessity common. The common substrate is connected to ground, but for purposes of simplicity in the accompanying schematics, the leads are not shown. It should also be noted that no arrowheads are shown on the MOST symbols. This is because integrated devices are symmetrical and no distinction is made between source and

drain. In fact, during circuit operation, some of the sources become drains and vice-versa.

Operation of RST Flip-Flop

The R S T flip-flop can operate as a shift register or a binary counter. When operating in the binary mode $\bar{Q}$ is connected to SET and Q is connected to RESET. The MOS transistors labeled A through K can be considered as switches which are closed when a negative voltage is applied to the gate (See Figure 12). The basic flip-flop consists of cross-coupled transistors F and G. Transistors E and H simply parallel F and G so that momentarily returning the gate of either E or H to the negative supply will set the flip-flop to a given state. Both gates must not be made negative simultaneously or an ambiguous state will result. In practice, both of these gates will be grounded until it is desired to reset the counter to zero, at which time one of the gates will be made negative for a short time.

After returning the DC Reset line back to zero volts, operation of the circuit is as follows: The clock signal is zero so that A is turned off. Hence, the drain is at $-V_s$. The gates of B and K are tied to the drain of A so that these transistors are turned on. Since $\bar{Q}$ (the drain of transistor C) is negative, C_1 is charged negative. Q is at zero so C_2 is uncharged. Even though C_1 is charged negative, thus placing a negative voltage on the gate of C, no current flows through the transistor because its source is connected to transistor D which is an open switch (it's an open switch because zero voltage at the clock is applied to the gate of D).

However, when the clock goes negative, transistor D is turned on. Since B is then turned off, the charge on C_1 does not leak off and the gate of C stays negative. C therefore stays on. Thus with C and D both turned on, current flows through the load resistor R_1. The regenerative feedback of the cross coupling then switches the state of the flip-flop, making Q negative and $\bar{Q}$ zero. The next cycle of the clock signal operates transistors I and J in an identical manner, returning the flip-flop to its original state. The flip-flop therefore goes through one complete cycle for each two cycles of the clock signal.

Use of capacitors C_1 and C_2 as storage elements places some restriction on the clock input waveform. This restriction is in the form of a minimum rise time of approximately 10 msec associated with the input signal. This is a very mild restriction.

Operation of the R S T flip-flop in the shift register mode is the same as for the binary, with one exception – instead of the SET and RESET lines being connected to $\bar{Q}$ and Q, they are connected to the datum and its complement. The datum shifts into the flip-flop when the clock is "0" and appears at the output when the clock goes to "1".

Operation of 21-Bit MOS Shift Register

The General Instrument MOGISTER$^{T.M.}$, a twenty-one bit shift register (MEM501) is really 3 registers in one package, each sharing common supply and clock voltages. The three can be used either independently or connected in series to give a total of 21 bits of delay to an arbitrary data stream.

Each bit of delay has a cross-coupled flip-flop to allow storage of data indefinitely between shift pulses. Only a single phase shift pulse has to be supplied; the additional 180° out-of-phase pulse is generated by an inverter in the chip.

The outputs will change on the trailing edge of the shift pulse, i.e., when ϕ goes from –10 to 0 volts. However, there is sufficient built-in delay so that an output won't start changing appreciably until the shift pulse is completely to zero (provided shift pulse fall time < 100 nsec). This delay makes it possible to transfer data from outputs of shift registers to inputs without any difficulty. In general, however, precautions should be taken to make sure data does not change during the shift pulse fall time.

The operation of the MFM501 is best understood by considering each of the transistors as a switch which is closed when a negative potential is placed on its gate, and open when its gate is at ground potential..

Operation begins by assuming some potential, say $-V_s$, which is applied to the input (see Figure 13). When the clock, ϕ, goes negative, Q_1 conducts charging capacitor C_1 and turning on Q_2. Q_6 is also turned on by ϕ which means $\bar{\phi}$ is at ground. With $\bar{\phi}$ at ground, Q_5 and Q_3 are off. With Q_3 off, C_2 is isolated except for leakage currents and retains its previous charge. Since C_2 is attached to the gate of Q_4, Q_4 also remains in its previous state. As ϕ goes to ground, Q_1 opens, isolating C_1 from the input. C_1 thus retains its negative charge and holds Q_2 on, which maintains ground potential on the drain of Q_2.

With ϕ at ground, $\bar{\phi}$ goes negative, turning on Q_3 and Q_5. As Q_3 turns on, C_2 is discharged to ground and Q_4 is turned off, making V_1 fall towards $-V_s$. With Q_5 conducting, the new negative potential V_1 is applied to the gate of Q_2, latching it in the on condition. (The time constant associated with the latch Q_5 is designed to be longer than that associated with Q_3 to insure no

race problem exists.) Thus after one complete cycle of the clock, the negative potential which was applied at the input appears at the output V_1.

Briefly then, the datum was shifted into the first stage while ϕ was negative and set Q_2 in the desired state. Interaction with Q_4 was prevented by opening switches Q_3 and Q_5. During this time, Q_4 "remembered" its state because C_2 was unable to discharge. (It is very important that Q_4 retain its state during the time ϕ is negative since it is precisely during this time that the output of Q_4 is being transferred to the input of stage 2). When ϕ went to ground, Q_2 remembered this state by virtue of the charge on C_1 and transferred this state (inverted) to Q_4. C_1 was required to hold its charge only during the transition time of the clock, since the closing of Q_3 and Q_5 latched the stage in its given state.

Because of the latching feature, MOGISTERS are capable of operation down to dc. However, the clock waveform must be asymmetrical because, as was noted above, C_1 needed to hold its charge only during the transition time of the clock, but C_2 was required to hold its charge during the entire time that the clock was negative. Even though the leakage current is only about 10^{-11} amperes, the capacitance is less than 1 pf, so the time constant is in the order of milliseconds. It is therefore recommended that the clock remain negative no longer than 50 microseconds; however, the clock can be at ground for any desired period of time. At frequencies below 10kcps, a simple one-shot multivibrator works well as the clock input. Above 10 kcps, a square wave generator is suitable.

Summary

The ability to fabricate a complete 21-bit shift register on a single chip constitutes a "giant step" forward in the art of monolithic circuitry and is a good indication that the MOS technology offers many possibilities beyond even this achievement. It is significant and typical of the power of this new technology that such a complicated circuit *can* be made to economically acceptable yields. This means that along with the promise of truly spectacular circuits in the months ahead, the age of the MOS microcircuit is here right now.

GENERAL INSTRUMENT CORPORATION
MICROELECTRONICS DIVISION
600 West John Street, Hicksville, L.I., N.Y. 11801